欧美建设发展科学数据中心的经验及对我国的启示

石 蕾 高孟绪 徐 波 王瑞丹
（国家科技基础条件平台中心，北京 100038）

摘要：大数据时代对科学数据的建设发展提出了新的挑战，世界各国科学数据中心快速发展，形成了一批具有较强国际影响力的科学数据中心，并开展数据资源管理与共享服务。在调研欧美等国家具有较强影响力的科学数据中心建设发展情况的基础上，总结数据中心在数据库建设、数据管理应用服务平台研发、数据全生命周期管理、数据服务等方面的特点，提出加强我国科学数据中心建设的思考与建议。

关键词：科学数据；数据管理；数据共享；科学数据中心；国际经验
DOI：10.3772/j.issn.1674−1544.2022.03.004 **CSTR：**15994.14.issn.1674.1544.2022.03.004
中图分类号：G311 **文献标识码：**A

Experience of Foreign Scientific Data Center Construction and Development and Its Enlightenment to China

SHI Lei, GAO Mengxu, XU Bo, WANG Ruidan
(National Science and Technology Infrastructure Center, Beijing 100038)

Abstract: The era of big data poses new challenges to the construction and development of scientific data. Scientific data centers around the world have developed rapidly, forming a number of scientific data centers with strong international influence on data management and sharing. This paper summarizes the characteristics experience of foreign scientific data centers in database construction, data sharing platform, data life-cycle management and data service, then puts forward some suggestions on strengthening the construction of scientific data centers in China.

Keywords: scientific data, data management, data sharing, scientific data center, international experience

0 引言

大数据的快速发展把科学研究带入以数据密集型科学研究为特点的“第四范式”，科学数据成为科研工作乃至国家发展的重要战略资源。科学数据是科技创新活动的重要产出，各类大型科研基础设施、科研观测网络建设运行以及科学实验等均产生了大量的科学数据。这些数据也成为新一轮科技创新活动的重要支撑。海量科学数据在形成过程中具有广泛分散性的特点，而通过数据的有效集成能够发挥其更大的价值。因此，世界各国积极推进科学数据中心建设，通过数据中

作者简介：石蕾（1982—），女，国家科技基础条件平台中心研究员，研究方向为科技资源管理；**高孟绪**（1982—），男，国家科技基础条件平台中心研究员，研究方向为科技资源管理；**徐波**（1988—），男，国家科技基础条件平台中心副研究员，研究方向为科技资源管理；**王瑞丹**（1965—）女，国家科技基础条件平台中心研究员，研究方向为科技资源管理（通信作者）。

基金项目：国家自然科学基金面上项目“基于动态与异构场景的科学数据中心评价方法研究”（7207041829）。
收稿时间：2022 年 2 月 10 日。

心开展科学数据的汇聚、管理、存储、开放与利用，其建设方式有自上而下的国家数据中心模式、自下而上的学科领域科学数据中心模式等[1]。

近年来，科学数据作为传播速度最快的科技资源，科学数据中心的建设发展越来越受到各方关注与重视，许多国家已将科学数据中心纳入本国重要的战略科技力量和重要的基础设施予以支持，形成了一大批具有较强影响力的科学数据中心。我国长期支持科学数据管理与开放共享工作，目前已在不同领域形成20个国家科学数据中心，在各政府部门、科研机构也形成了一批层次不同、类型多样的科学数据中心，为推动科学数据共享共用、提高资源利用效率发挥了积极作用。但是由于我国建议科学数据中心起步较晚，建设运行机制尚不健全等问题依然突出，与欧美等国家已建成的科学数据中心相比，仍然存在系统性的差距和不足。因此，本文将梳理和总结欧美等发达国家在建设发展科学数据中心方面的经验和做法，为我国建设发展国家科学数据中心提供参考借鉴。

1　欧美建设发展科学数据中心的主要做法

欧美等许多发达国家很早就注重对科学数据进行积累、有效管理与长期保存，依托科研机构或高校陆续建设了若干国家级科学数据中心，依托科学数据中心开展相关学术领域科学数据汇聚，面向本国及全球开放共享，对国家的科学技术、教育与国民经济发展发挥重要的作用。同时，欧美等发达国家通过制定国家政策支持科学数据的管理与共享，依托科学数据中心汇聚整合各类科学数据，建立了适合科学数据中心发展的管理机制，形成了有益于科研活动的数据生态[2]。近年来，笔者对欧美等发达国家建设发展科学数据中心进行了调研，从数据资源建设、数据资源管理、基础设施建设、数据中心人才队伍建设、可持续发展5个方面归纳总结了其成功的经验和做法。

1.1　在数据资源建设方面，注重高质量、长序列科学数据库建设

科学数据中心以科学数据为主要管理对象，存储及可使用数据的数量和质量是科学数据中心能力建设和发展最重要的因素。科学数据中心十分重视科学数据的整合范围、数据质量以及对历史数据的整理与汇集，各数据中心都在积极建设领域内完整、权威且高质量的科学数据库，将建设数据丰富、内容完整、信息准确的科学数据库作为科学数据中心建设的重要内容，以此形成科学数据中心的核心优势。

如在材料科学领域，数据库已成为材料基因工程的重要组成部分。由德国波恩大学于1913年创建了ICSD无机晶体结构数据库，通过广泛整合依托高质量期刊出版的无机晶体结构详细信息，建成涵盖金属、合金、陶瓷等非有机化合物的晶体结构数据库，整合20余万种晶体结构数据，已成为世界最大的无机晶体结构数据库，被材料领域科研人员广泛使用[3]。

再如在生命科学领域，欧美国家较早就启动建设核酸序列数据库。美国在1988年就关注到生物技术领域的重要性并成立了美国国家生物信息中心[4]，支持GenBank等数据库的建设并长期维护更新。美国国家生物信息中心NCBI通过与欧洲生物信息研究所EBI和日本DNA数据库DDBJ共同组建国际核酸序列数据库合作组织，依托其建立的为核酸序列数据分配唯一标识的机制，支撑其占领领域数据高地，通过机制建设促使全球数据持续向其汇聚，形成了具有较强影响力的核酸序列数据库。

1.2　在数据资源管理方面，围绕全生命周期开展科学数据管理

数据本身具有涉及面广、传播速度快等特点。随着网络化和智能化的发展，世界各国科学数据中心都将吸纳全球数据和数据服务全球作为数据中心建设发展的重要目标，并在数据管理政策中强调与国际相关法律条款和标准规范的一致性。为增强对科学数据的整合汇聚和服务能力，科学数据中心普遍开展数据的全生命周期管理，覆盖科学数据生产、处理、分析、保存、访问、重用等环节。

美国地球观测数据信息系统EOSDIS是美国

航空航天局NASA支持建设的综合地球观测数据管理和服务平台，旨在建立有利于数据充分利用和长期服务的数据共享系统[5]。其突出特点是建成了一体化的数据网络体系，形成多方共建、协调统一的数据互联互通机制，以及统一的基础设施体系，有效支撑了多学科综合性研究，支撑了对地球系统变化的理解和认知。美国地球观测系统数据信息系统（EOSDIS）是其下设各分布式数据存档中心的数据管理系统，承担数据的获取、保存、处理、分发，负责信息管理、网络建设、算法交换、产品发布等功能，支撑汇总海量地球观测数据产品、辅助数据和元数据[6]。EOSDIS通过统一的系统平台长期开展地表、生物圈、固体地球、大气、海洋等全球观测数据管理与开放共享，其搜索范围涉及数以百万计的文件和PB级数据，数据来源与世界各国的多格式数据[7]。

美国的国际地球科学信息网络中心（CIESIN）开展在线数据管理与空间数据集成，在世界范围内开展地球科学数据的收集、存储、归档、维护和共享，面向全球用户提供多种方式的数据浏览、在线分析和数据下载服务[8]。加拿大天文数据中心CADC提供加拿大—法国—夏威夷望远镜（CFHT）等天文观测数据服务，提供数据存储、共享、在线处理等全流程服务[9]。

1.3　在基础设施建设方面，积极开展数据管理与应用软件系统平台研发

完备的科学数据管理与应用服务平台是科学数据中心开展科学数据管理的重要基础设施，并研发与之相匹配的各类软件工作，以提高科学数据收集过程中的传输、编目、检索、分析等不同阶段的数据管理及使用需求。大数据技术对数据管理应用系统平台提出了更高的要求，推动了科学数据系统平台持续向支持海量、复杂数据的高速处理发展。各数据中心都在持续开展各类系统平台及软件工具的研发与更新，并对硬件平台进行升级与扩展，以满足对大规模、多类型数据的高效管理与分析挖掘。

美国国家生物信息中心管理并运行着全球影响力最高的生物医学领域科学数据中心，其软件平台以自主研发为主[10]，提供一系列数据检索、数据对比、进化树、分析结构分析等复杂生物信息的分析解决方案，并提供相应的方法学培训课程[11]。通过创建自动化系统来存储和分析有关生物学、生物化学和遗传学信息，序列比对软件BLAST已成为生命科学领域使用最多的数据和工具资源，通过序列相似性对比，可支持识别基因和遗传特征。其跨库搜索和检索系统Enterz可为用户提供对比序列、映射、分类和结构数据的集成访问。

在法国教育研究部于2012年发布的《2012—2020年研究基础设施国家战略》中，法国斯特拉斯堡天文数据中心（CDS）被称为“研究基础设施”[12]，其建设的天文数据库SIMBAD[13]是世界知名天体参考数据库。法国斯特拉斯堡天文数据中心（CDS）致力于天文数据和相关信息的收集和全球分发，努力将数据中心打造为“处于国际合作枢纽地位的一个数据中心”[12]，其建成的数据整合工具Aladin[15]是一个集访问、可视化和天文图像分析以及数据库及相关数据一体化交互的接口系统，有效地提高了全球天文数据的互操作能力和开放服务水平。

1.4　在人才队伍建设方面，强化专业化人才培养

科学数据工作涉及领域多、专业性强，要建立稳定高效的科学数据收集与管理体系和高质量的科学数据分析应用平台，人才队伍建设尤为重要。欧美等国科学数据中心除少量NCBI等规模相对较大外，其他工作团队普遍总体规模不大，但高水平的科研人员和专业化从事数据整理、归档、分析等方面的工作人员在工作团队中占比较大，而管理与辅助人员较少是其普遍特点。

如法国斯特拉斯堡天文数据中心（CDS）现有工作人员约40人，其中主要为固定人员，由天文科学家、软件工程师、档案专家及若干管理人员组成[16]。美国国家生物技术信息中心（NCBI）是规模较大的数据中心，拥有一个由计算机科学家、分子生物学家、数学家、生物化学家、研究医师和结构生物学家组成的多学科研究

小组，构建了多领域科研人员共同建设、管理和应用科学数据的机制，有效地带动了基于科学数据的交叉研究，有利于其开展高水平的计算分子生物学的基础和应用研究[17]。

德国地球与环境科学数据出版平台PANGAEA工作团队约50人，约半数人员从事数据编辑、管理与咨询服务工作，而很多长期参与数据中心工作的科研人员分别来自各专业研究团队[18]。德国地球科学领域数据中心GEOROC团队规模不大，由数据输入和系统管理相关人员构成[19]。

1.5　在可持续发展方面，数据中心建设与数据服务紧跟学术领域的发展

随着数据密集型科学研究范式的到来和快速发展，越来越多的科研人员在科研工作中注重科学数据积累并加强对科学数据的分析与应用，科学数据在学术领域发展中的重要性日益凸显。科学数据既是科技创新的重要基础，也是科技创新的重要产出。越来越多的政府科技管理部门、学术出版机构、国际科技组织等开展科学数据工作，并且多个国内外知名出版集团发布了明确的学术期刊相关科学数据的汇交与开放政策，专注于科学数据出版的期刊快速发展并日趋成熟，还有多个国际组织积极号召开放与共享科学数据。科技管理机构、学术机构等加强与各领域科学数据中心的联合与合作，使其成为科学数据积累和数据中心发展重要推动力的同时，也促使科学数据活动更多地融入了相关学术领域。

近年来，随着全球对科学数据的广泛重视，学术期刊将科学数据纳入其视野，与科学数据中心协同发展的态势尤为明显，进展也尤为迅速。如Spring Nature等生物医学领域国家主流学术期刊在接收论文的同时，也要求论文递交者把论文关联的序列数据递交到生物领域数据中心。与期刊的广泛合作，论文科学数据的汇交机制极大地促进了全球数据的汇集，筑牢了科学数据中心的数据资源基础，也提升了数据中心的全球服务能力。随着《Scientific data》《Biodiversity Data Journal》《Earth System Science Data》等数据出版期刊的快速发展，一批基于数据出版的科学数据仓储系统和科学数据中心逐步发展起来。

再如，德国地球与环境科学数据出版平台PANGAEA是一个对全球任何组织和个人开放并保证长期运行的地球科学数据库[20]，旨在归档、发布和分发地球系统研究的相关数据。与其紧耦合的数据期刊《地球系统科学数据》(《Earth System Science Data》)影响因子高，与期刊的紧密结合机制快速提升了数据中心影响力。法国斯特拉斯天文数据中心CDS通过支持创立国际虚拟天文数据台联盟，提升数据中心数字化水平和全球服务能力。

2　我国科学数据中心建设发展现状

近年来，随着我国科技创新投入的持续增加，大型科学装置建设运行、传感器和传感网络在科研活动中广泛应用，重大科学实验在多个领域系统开展，产生了海量科学数据，将我国科技创新活动也带入了以数据积累和应用为重要科研方式的新阶段。我国积极支持科学数据中心建设，广泛开展不同层面的科学数据共享平台建设与运行工作，形成了一批层次不同、类型多样的科学数据中心，为推动科学数据共享共用、提高资源利用效率发挥了积极作用。

2018年，国务院办公厅印发《科学数据管理办法》，明确提出在条件好、资源优势明显的科学数据中心基础上优化整合形成国家科学数据中心[21]。2019年，科技部、财政部组建首批20个国家科学数据中心，主要分布在地学、生命科学、基础科学等领域，我国科学数据工作进入新阶段。国家科学数据中心按照学科领域开展数据资源体系建设，持续开展科学数据资源的汇聚与长期保存，建设研发数据管理与共享服务平台和各类软件工具以提升数据资源的综合集成与治理能力，提高对科技创新活动在数据服务能力，并面向国家重大发起站战略、科技创新热点以及新冠肺炎疫情防控的民生发展需求开展数据资源服务，国家科学数据中心影响力稳步提升。

国家微生物科学数据中心依托单位并承建世界微生物数据中心，建设维护了微生物资源相关的系列重要数据库，包括全球微生物保藏机构数据库、全球微生物菌种资源目录、全球微生物参考菌株数据库、微生物资源引用数据库等[22]。我国作为基因组数据产出大国，生物组学数据量约战全球的40%[23]。国家基因组科学数据中心建设的GSA数据库面向全球开展组学数据的汇交、存储、管理与共享，已成为国际主要生物数据库之一。建设的GSA−Human人类遗传资源数据库，支持类型数据汇聚与管理，有效支撑了我国人类遗传资源数据的安全管理与开放共享，服务于国家面向人口健康和生命安全相关的科研活动[24]。

国家对地观测科学数据中心建成了国内规模最大的对地观测科学数据共享资源库，数据资源覆盖我国所有国家级卫星数据和规模以上商业卫星公司[25]。其建成的国家综合地球观测数据共享平台面向地球观测组织GEO开展数据共享和应用。国家天文科学数据中心建设的郭守敬望远镜LAMOST数据发布系统，已成为基于LAMOST千万量级的光谱数据开展银河系结构、形成和演化研究的重要基础。

3　对我国科学数据中心建设的建议

随着物联网、5G和人工智能为代表的信息技术持续飞速发展，全球科技活动产生的数据从宏观到微观急剧增长，将对科学数据中心提出更高的要求，也必将对我国科学数据管理机构适应并引领科研需求提出新的挑战。与世界发达国家具有国际影响力的科学数据中心相比，我国科学数据中心普遍存在总体实力相对较弱，高质量、高影响力的科学数据库相对较少，科学数据管理与分析应用平台水平不高，数据中心专业化人才缺乏等问题，在多渠道整合科学数据、形成科学数据中心核心优势的手段不多，与本领域科学共同体和相关机构的合作较少，多方利益共赢方面的机制不够健全。借鉴发达国家科学数据中心建设发展的经验做法，建议在以下几方面持续加强科学数据中心建设。

3.1　系统谋划科学数据中心数据积累、管理分析平台研发与基础设施建设

科学数据资源、数据管理与分析应用服务平台，以及承载数据存储应用等工作的科研基础设施是科学数据中心建设的核心要素，三者缺一不可，必须并行发展。科学数据资源是数据中心建设基础，科学数据资源的数量和质量关乎科学数据中心的生命力；数据管理与分析平台是数据资源汇聚、管理、开放、应用全生命周期的中枢系统，直接决定着数据中心的数据管理与应用服务水平，是发挥数据资源价值的核心系统；数据基础设施主要是指数据中心运行所需机房、网络等硬件条件，是数据中心建设运行的条件保障。这三者在与数据中心的运行管理制度和机制合理配合的情况下，构成了高水平科学数据中心建设的基本条件。

科学数据中心应在深入分析我国科学数据资源现状和相关学科领域发展趋势的基础上，认真研究并系统规划国家科学数据中心发展目标和发展思路，梳理完善各领域科学数据资源体系，对接我国科技创新规划和重大创新需求，突出重点建设一批高质量科学数据库，建立完善科学合理的数据汇聚与管理规范。以全球视角采取更加开放的态度谋划国家科学数据中心开放应用服务平台建设发展，面向全球用户提供高质量的科学数据开放服务。

3.2　开展高质量数据库建设，形成典型优质的科学数据产品

加强战略性、基础性科学数据库建设，紧密对接相关领域科研活动实际需求，提升科学数据库质量，建立稳定的科学数据汇聚渠道和高水平的数据质量控制机制，形成持续更新的且满足科研需求的高质量科学数据库。尤其是要面向流通需求形成科学数据产品，持续满足科研人员对便捷获取科学数据资源的广泛需求。可基于同一科学数据库，针对科学研究、区域发展、企业不同的应用需求，形成多样化的科学数据产品。如面向科研的需求，可针对国家科技战略部署，针对某一研究方向中创新链的不同环节，研发系列

科学数据产品，支持不同研究团队开展创新研究。再如面对区域发展的需求，可在抽取研究区域数据以及邻近区域或相似区域科学数据的基础上，研发面向区域发展布局的科学数据产品。又如面向企业创新的需求，可根据企业创新研发需求，结合产业链上下游研发数据产品，支撑创新发展。

3.3　完善基于全生命周期的科学数据管理机制，增进学术融合

围绕科学数据形成、成长、成熟、衰亡的生命基本过程，将科学数据管理贯穿科学数据生命周期，包括数据收集、数据认证、数据加工、数据保存、数据发布、数据共享及数据处置等各个环节。以科学数据生命周期为主要轨迹，加强相关政策制度制定，完善运行管理机制，优化标准体系。将科学数据管理融入科研活动生命周期，以科技计划项目科学数据汇交为切入点，优先加强政府预算资金资助的科技计划项目形成的科学数据的全生命周期管理，建立科研人员生产数据并向数据中心汇交，科学数据中心开展数据整理、保存及数据服务，并将数据开放共享情况反馈科研管理机构及科研人员，形成闭环管理。积极推动学术论文相关科学数据管理与共享，促进科研论文相关科学数据向数据中心汇交，在促进科学数据中心发展的同时，让科学数据在科研支撑、学术传播和科研诚信体系建设等方面发挥更大作用。

3.4　打造有效服务科研人员的科学数据管理与应用平台

面对信息技术的快速发展，需要深刻理解大数据对科学研究思维模式和研究范式带来的影响，以及大数据技术对科学数据管理与分析应用技术革新的影响，加强科学数据管理与应用的科研平台建设，开展科学数据整合与分析挖掘软件工具研发，打造科研领域的数据分析应用服务平台。在现有数据资源的基础上，根据科学数据全生命周期建立完整的科学数据管理应用工作流程和数据平台技术流程，打通系统平台中心数据提交、质控、整理、编目、存储、应用各环节，完善系统平台对异构数据的兼容能力和多指标数据的识别和汇聚能力，提升数据平台对大规模、复杂性数据的实时处理和智能发现能力，完善基于多用户的数据应用服务支撑系统。开展科学数据管理应用相关软件工具研发，开发智能化科学数据接收与质量审核软件工具，研发科学数据分析挖掘方法、算法、模型。根据科研人员、科研团队、科研机构等科学数据管理需求，完善数据平台相关标准规范与流程，建立完善面向复杂事件的科学数据管理分析应用服务平台，打造满足科研工作需求的科学数据研究平台。

3.5　加强国家科学数据中心建设，重视专业化的人才培养

充分发挥国家科学数据中心的作用，做好相关领域科学数据的汇聚整合、存储管理与开放应用。围绕科学数据全生命周期，建立健全各领域科学数据标准体系，在规范数据中心科学数据管理的基础上，逐步形成全社会科学数据管理标准化、规范化共识，促进科学数据资源质量提升。建立完善科学数据资源标识体系，对汇入国家科学数据中心的科学数据资源进行统一标识，支持科学数据资源的可定位、可访问、可确权。提升国家科学数据中心全球服务能力，加强与世界各国科学数据中心的交流与合作。探索建立适合国家科学数据中心建设发展的人员培养与晋升机制，培养专业化、复合型的科学数据应用服务人才，优化适合稳定高水平人才的人员晋升发展机制，探索保障科研人员贡献的数据开放共享权益保护新机制。建立稳定的基础设施更新完善机制，形成绿色节能可持续的数据运行环境。

参考文献

[1] 王明明, 王卷乐, 赵强, 等.ICPSR科学数据中心的建设经验与启示[J].中国科技资源导刊, 2017(6): 100−107.

[2] 黄铭瑞, 李国庆, 李静, 等.国家科学数据中心管理模式的国际对比研究[J].农业大数据学报, 2019(4): 14−28.

（下转第110页）